Supplementary Note

alpine fragment sequence bias model

For each sample and each transcript (each isoform of a gene), counts $(0, 1, 2, \dots)$ of aligned paired-end fragments are tabulated based on the starting position p of the read closest to the start of the transcript and the length of the fragment l into a matrix M. The following model applies to paired-end RNA-seq fragments, but could be modified for single-end data by inferring approximate fragment lengths. Fragments with start and end positions which were not observed will have $M_{pl} = 0$. For computational efficiency, the range of l is limited to the middle ∼99% of the empirical fragment length distribution. For the IVT-seq samples, the center of the distribution was defined by $l \in [100, 350]$ and for the GEUVADIS samples, $l \in [80, 350]$, with a total of L possible fragment length values considered. Note that most entries of M will be 0. For estimating bias parameters and evaluation of predicted coverage, the fragments which begin on the first basepair or end on the last basepair of a transcript are not included, as large counts for these potential fragments could impair estimation of the coefficients for coverage biases within the body of the transcript. The matrix M then represents nearly all of the potential fragment types that could arise from a given transcript. Paired-end reads which are compatible with multiple isoforms are assigned to each of the transcript matrices M, as long as the fragment length is within the range defined above. The counts in the matrix are modeled on a number of features including:

- The fragment's length, l.
- The relative position of the fragment in the transcript.
- The GC content of the fragment (and other features of the fragment sequence).
- The sequence in a 21 bp window around the starts of the two paired-end reads. The *Cufflinks* variable length Markov model (VLMM) for random hexamer priming bias described by Roberts et al. [1] is used.

For notational simplicity in the following sections, the counts in the matrix M are collapsed into a vector Y indexed by j such that Y_j is a count for the j-th potential fragment type.

Estimating bias offsets and coefficients

Bias terms are estimated using a subset of the genes that have only one isoform, which avoids the task of probabilistic assignment of the fragments from different isoforms of a gene. The bias terms were estimated across the 64 transcripts in the IVT-seq dataset (in two batches for cross-validation), across 100 medium to highly expressed genes in the GEUVADIS dataset, and across 200 genes for the SEQC and ABRF datasets. In this section, we distinguish model terms into two groups: bias *offsets*, which are calculated independently of other terms, and bias *coefficients* which are estimated within a Poisson generalized linear model (GLM) described below. The offsets are included as fixed terms in the GLM when estimating the coefficients. For calculating all bias terms, the fragments

from highly expressed genes are down-sampled, so that each gene contributes nearly equally according to an initial estimate of FPKM (without bias correction) to the model terms. This is similar to a weighting procedure for bias parameter estimation used by Roberts et al. [1].

Two offsets are calculated using the fragments that align to genes with a single isoform. The first is an offset for the fragment's length (FL). An empirical distribution of fragment lengths is calculated by examining the lengths of fragments aligning to the given subset of genes. The log of the kernel density estimate of fragment lengths evaluated at each fragment length l defines the fragment length offset.

The second offset is for read start sequence preferences (RS) caused by differential binding efficiencies of random hexamer primers [2]. For this, *alpine* includes an implementation of the 21 bp VLMM for read starts proposed by Roberts et al. [1] and used in the bias correction method of *Cufflinks*. The model defines the probability of observing a read aligning to a position, given the sequence context surrounding the read start position. The VLMM is defined by the following sequence: four 0-order positions, followed by three 1-order positions, ten 2-order positions, two 1-order positions, and finally two 0-order positions. A 0-order position corresponds to frequencies for the four nucleotides at the given position, a 1-order position to frequencies for di-nucleotides ending on the given position, and so on. The read start position itself is the second of the ten 2-order positions (for a diagram, see Supplementary Methods Figure 2 of Roberts et al. [1]). Observed frequencies are collected by examining the position-specific nucleotide, di-nucleotide and tri-nucleotide frequencies surrounding the read starts of fragments aligning to the given subset of genes (5' and 3' reads tabulated separately). Expected frequencies are estimated from the given subset of genes assuming a uniform distribution of fragments to all positions (5' and 3' separately). The read start bias offset for a given fragment type j is the log of the observed over expected probabilities according to the VLMM for 5' and 3' read start positions added together. Positions without sufficient sequence context are not included in the observed frequency calculations, and a truncated version of the VLMM is used for calculating the bias offset of these potential fragments.

A number of coefficients are then included in a GLM. These coefficients include: a natural cubic spline for the relative position (RP) of the fragment in the transcript with knots at $0.25, 0.5, 0.75$ and boundary knots at $0, 1$; a natural cubic spline based on the GC content for each fragment (GC) with knots at $0.4, 0.5, 0.6$ and boundary knots at $0, 1$; and four indicator variables that indicate if the fragment contains a stretch (STR) of higher than 80% or 90% GC content in a 20 or 40 bp sequence within the fragment.

These variables are used to construct a GLM that accounts for the biases described above, specifically to model the count Y_j of fragment j as:

$$Y_j \sim \text{Poisson}(\lambda_j^b)$$

$$\log(\lambda_j^b) = \sum_k X_{jk}\beta_k + o_j + g_j$$

where k indexes the columns in the design matrix X and β_k is the matching coefficient. The matrix X contains the spline basis vectors for relative position and fragment GC content, and the four indicator variables for GC stretches. The o_j term represents the offset

for the fragment length and optionally the read start bias for fragment j. The g_j term represents any baseline differences in gene expression for the gene that fragment j aligns to. For fitting the bias model to a collection of fragments across multiple genes, any differences in gene abundance are nuisance parameters, which can be accounted for by adding gene-level expression terms such as g_j, or by down-sampling fragments such that all genes have nearly equal initial FPKM estimates. Here we both use the down-sampling strategy described earlier for calculating bias offsets, and any residual differences in gene-level expression are removed by including the term g_j.

The estimated coefficients for relative position, GC content, and GC stretches are obtained using the *splines* and *speedglm* R packages. We do not model an interaction between fragment length and GC content of the fragment, as discussed by Benjamini and Speed [3], because we were able to predict coverage drops with GC content alone and so to limit the number of parameters in the model. However, such an interaction could be added at this stage.

Predicted read start coverage

The predicted read start coverage for position p is defined as:

$$\widehat{v}_p = \sum_{j:\, c(j)=p} \hat{\lambda}_j^b$$

where $c(j) = p$ indicates that the j-th potential fragment covers position p. Note that the predicted coverage $\widehat{v}_p$ is only used for plotting and model comparison on the IVT-seq dataset, and not for estimation of the coefficients or transcript abundance. For estimation of the coefficients and transcript abundance, only the fragment-level counts Y_j and estimates $\hat{\lambda}_j^b$ are used.

alpine bias models per dataset

To concisely describe the bias models used, we introduce the following acronyms. For description and details on the estimation of these terms, see the preceding section, "Estimating bias offsets and coefficients".

RS	read start VLMM [1]
FL	fragment length
RP	relative position
GC	fragment GC content
STR	fragment GC stretches

The following terms were included in the bias models used to analyze the various datasets. These terms were fitted using the model described above, where the inclusion of multiple terms implies they have an additive effect on the log of the rate of fragment counts. The read start and fragment length terms are pre-calculated as described above and included as offsets.

dataset	model	RS	FL	RP	GC	STR
IVT	GC		X		X	
IVT	GC+str		X		X	X
IVT	read start	X	X			
IVT	all	X	X		X	X
GEU	GC+str		X	X	X	X
GEU	read start	X	X	X		
GEU	all	X	X	X	X	X
Sim	GC		X		X	
SEQC	GC		X	X	X	
ABRF	GC		X	X	X	

In the IVT-seq dataset, relative position (RP) was not included, as strong positional bias was not observed, as opposed to the poly-A selected GEUVADIS, SEQC, and ABRF datasets, which did exhibit positional bias.

Transcript abundance estimation

The estimated bias terms for a given sample can be used to improve the estimates of transcript abundance for that sample for genes with a single isoform or multiple isoforms. For fragment type j and isoform i, the Poisson fragment rate is given by:

$$\lambda_{ij} = \theta_i a_{ij}$$

where A is a sampling rate matrix, and $\vec{\theta}$ represents the abundance of the different isoforms of a gene, as described by Salzman et al. [4] and Jiang and Salzman [5]. $a_{ij} = 0$ if fragment type j could not arise from isoform i. If fragment type j can arise from isoform i, one possible parametrization sets $a_{ij} = q(l_j)N$, where q is the empirical density of fragment lengths, l_j is the fragment length of the j-th fragment type, and N is the total number of fragments which are compatible with the annotated genes. Here, we included the fragment length in the overall bias term $\hat{\lambda}^b_{ij}$, so we set $a_{ij} = \hat{\lambda}^b_{ij} N/(L \cdot 10^9)$ when fragment type j can arise from isoform i. The denominator contains 10^9 and the range of the fragment lengths L considered in the model, so that final estimates of $\hat{\theta}$ are on the FPKM scale.

$\hat{\lambda}^b_{ij}$ can be calculated using the estimated coefficients $\hat{\beta}_k$ and the estimated offsets $\hat{o}_j$:

$$\log(\hat{\lambda}^b_{ij}) = \sum_k X_{jk}\hat{\beta}_k + \hat{o}_j$$

Because we found that $\hat{\lambda}^b_{ij}$ was identical across isoforms for nearly all fragments j, in this paper to simplify the model we approximated λ^b_{ij} with λ^b_j. Note that $\sum_i \hat{\theta}_i$ is not equal to 1, as these represent expression abundances, so they are only required to be non-negative. The Poisson model defined here is the same model as proposed by Jiang and Salzman [5], but here we explicitly model the bias using offsets $\vec{o}$, a matrix of features X, and a vector of log fold changes $\vec{\beta}$, so not the same β as defined by Jiang and Salzman [5].

The log likelihood for a given estimate $\hat{\theta}$ of the isoform abundances is evaluated by:

$$\ell(\hat{\theta}|A) = \sum_j \log f(Y_j, \sum_i \lambda_{ij})$$

where f is the probability density function for a Poisson random variable. The maximum likelihood estimate of $\hat{\theta}$ is obtained using an EM algorithm as described by Jiang and Salzman [5], where fragment types j with no observed fragments need only be considered for certain steps of the EM.

For genes with a single isoform, the maximum likelihood estimate for θ, given the estimated bias terms $\hat{\lambda}_j^b$, is:

$$\hat{\theta} = \frac{L \cdot 10^9 \sum_j Y_j}{N \sum_j \hat{\lambda}_j^b}$$

As the bias terms associated with GC content and relative position curves introduce an arbitrary intercept, the estimates $\hat{\theta}$ for all transcripts for a sample are multiplied by a scaling factor, $\bar{\lambda}$. This scaling factor is the average over all T transcripts of the average bias term over all J_t fragment types within transcript t.

$$\bar{\lambda} \equiv \frac{1}{T} \sum_{t=1}^{T} \left[\frac{1}{J_t} \sum_{j_t=1}^{J_t} \hat{\lambda}_{j_t}^b \right]$$

For normalizing transcript abundances across samples, the $\hat{\theta}$ estimates are scaled using the median-ratio method of *DESeq* [6].

Transcript quantification for GEUVADIS, simulated, SEQC, and ABRF datasets

Cufflinks version 2.2.1 [1, 7] was run with bias correction turned on, with the commands:

```
cuffquant -p 40 -b genome -o cufflinks/file genes.gtf \
  tophat/file/accepted_hits.bam
cuffnorm genes.gtf -o cufflinks cufflinks/file1/abundances.cxb \
  cufflinks/file2/abundances.cxb ...
```

RSEM version 1.2.11 [8] was run with the commands:

```
rsem-prepare-reference --gtf genes.gtf genome rsem/hg19
rsem-calculate-expression -p 20 --no-bam-output --paired-end \
  <(zcat fastq/file_1.fastq.gz) <(zcat fastq/file_2.fastq.gz) \
  rsem/hg19 rsem/file/file
```

For the ABRF dataset, *RSEM* was run with the additional flag `--estimate-rspd`, as this dataset exhibited differential positional bias across protocol. For the simulated data, the flag `--no-polyA` for `rsem-prepare-reference` was used.

kallisto version 0.42.4 [9] was run with bias correction turned on. The `transcripts.fa` file was generated using `rsem-prepare-reference`. The following commands were used:

```
kallisto index -i kallisto_index transcripts.fa
kallisto quant --bias -i kallisto_index -o kallisto/file \
  fastq/file_1.fasta fastq/file_2.fastq
```

Salmon version 0.6.0 [10] was run with bias correction turned on. The `transcripts.fa` file was generated using `rsem-prepare-reference`. The following commands were used:

```
salmon index -t transcripts.fa -i salmon_index
salmon quant -p 10 --biasCorrect -i salmon_index -l IU \
  -1 fastq/file_1.fastq -2 fastq/file_2.fastq -o salmon/file
```

For the ABRF dataset, *Salmon* was run with the additional flag `--useFSPD`, as this dataset exhibited differential positional bias across protocol.

Sailfish version 0.9.0 [11] was run with bias correction turned on. The `transcripts.fa` file was generated using `rsem-prepare-reference`. The following commands were used:

```
sailfish index -t transcripts.fa -o sailfish_index
sailfish quant -p 10 --biasCorrect -i sailfish_index -l IU \
  -1 fastq/file_1.fastq -2 fastq/file_2.fastq -o sailfish/file
```

MISO version 0.5.3 was run on the GEUVADIS dataset using STAR aligned reads with the following commands, assuming read of length 75 bp and a fragment length distribution centered at 160 bp with a standard deviation of 30 bp:

```
index_gff --index genes.gtf miso_index
miso --run miso_index miso/file.bam --output-dir miso/file --read-len 75 \
  --paired-end 160 30 -p 10
```

References

[1] Adam Roberts, Cole Trapnell, Julie Donaghey, John Rinn, and Lior Pachter. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol*, 12(3): R22–14, 2011.

[2] K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*, 38(12):gkq224–e131, 2010.

[3] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, 40(10):e72, 2012.

[4] Julia Salzman, Hui Jiang, and Wing H. Wong. Statistical modeling of RNA-seq data. *Stat Sci*, 26(1):62–83, 2011.

[5] Hui Jiang and Julia Salzman. A penalized likelihood approach for robust estimation of isoform expression. *Stat Interface*, 8(4):437–445, 2015.

[6] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106+, 2010.

[7] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, 2010.

[8] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4): 493–500, 2010.

[9] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, 34(5):525–527, 2016.

[10] Rob Patro, Geet Duggal, and Carl Kingsford. Accurate, fast, and model-aware transcript expression quantification with salmon. *bioRxiv*, pages 021592+, 2015.

[11] Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*, 32(5):462–464, 2014.